Párhuzamos igei szerkezetek közvetlen kinyerése párhuzamos korpuszból

Sass Bálint

MTA Nyelvtudományi Intézet
e-mail: sass.balint@nytud.hu

Kivonat Jelen dolgozatban egy egynyelvű korpuszra kifejlesztett, igei szerkezeteket kinyerő eljárást alkalmazunk holland-francia párhuzamos korpuszra, a korpuszreprezentáció alkalmas átalakításával, kétnyelvű, párhuzamos igei szerkezetek kinyerése céljából. Az igei szerkezetek közül kiemelendők a vonzatos komplex igék, melyekben az ige mellett egy névszói kollokátum, valamint vonzat is áll (pl.: *részt vesz vmiben*). A nyelvtechnológiai alkalmazások (pl.: a gépi fordítás) lexikális erőforrásainak tartalmaznia, ismernie kell ezeket a kifejezéseket, hogy magas nyelvi minőségű kimenetet adhassanak. Ezek a szerkezetek ugyanakkor sok esetben más nyelvre lefordítva teljesen más formát mutatnak. Bár a szükséges elemző lépések során alkalmazott egyszerű közelítő módszerek, valamint a feladat nehézsége miatt a kinyerés pontossága nem kiemelkedő, jelen dolgozatból világos, hogy az algoritmus képes különféle, akár aszimmetrikus, párhuzamos szerkezetek feltérképezésére is.

1. Bevezetés

A többszavas kifejezésekkel foglalkozó szakirodalom legnagyobb része a kételemű, két tagból álló kifejezésekkel foglalkozik [1]. Siepmann [2, 412. oldal] szerint általánosan elfogadott a kutatók között, hogy a kollokációk bináris egységek. Számtalan asszociációs mértéket dolgoztak ki [3], melyekkel két tag közötti kapcsolat szorossága mérhető. A kettőnél több tagú kifejezések kezelésével ritkábban foglalkoznak, az ide tartozó módszerek három csoportra oszthatók [4, 5.1 fejezet]: egyrészt megpróbálhatjuk az asszociációs mértékeket kettőnél több elemre kiterjeszteni; alkalmazhatunk iteratív kollokációkinyerő módszereket, ahol a már kinyert kéttagú kollokációk a következő iterációban összevont elemként egy nagyobb kiterjedésű kollokáció részét képezhetik; valamint a kinyert bigramokat utólag feldolgozva is következtethetünk bizonyos többtagú kollokációk meglétére.

A többelemű kifejezések között speciális csoportot alkotnak a vonzattal is bíró komplex igék. Ide tartozik a *kilátásba helyez vmit* vagy a *részt vesz vmiben*. Ezekben a szerkezetekben *négy* egységet különíthetünk el: az igét, a vonzatot (magyarban esetrag képviseli), a komplex ige névszói elemét, valamint e névszói elem esetragját. A nyelvekre általában jellemző, hogy a komplex igék névszói elemét és a vonzatot *ugyanazokkal* a nyelvi eszközökkel kapcsolják az igéhez, legyen az esetrag, névutó, elöljáró, igei partikula vagy akár sorrendi megkötés, mint az angol tárgy esetében. Emiatt ezek a „négyelemű kollokációk” speciális

kezelést igényelnek: az őket megcélzó lexikai kinyerő eljárásnak fel kell ismernie, hogy az adott bővítményi elem lexikálisan kötött módon a komplex ige része-e (*kilátásba*, *részt*), vagy pedig vonzat, mely esetben a konkrét szó nem része a szerkezetnek, csupán a viszonyjelölő (*vmit*, *vmiben*).

Nyilvánvalónak tűnik, hogy ezek a szerkezetek csak a vonzatukkal együtt teljesek, csak teljes formájukban tudnak hozzájárulni nyelvtechnológiai alkalmazások teljesítményének javításához, például tipikusan egy gépi fordítóban használt lexikai adatbázis elemeként. Mégis a korábbi kutatásokra jellemző, hogy elfogadják helyes eredménynek a hiányos szerkezeteket is. A kollokációkutatók sokszor megfeledkeztek arról, hogy a kollokációknak vonzatuk is lehet, amint ez az [5] cikkben idézett *zur Verfügung stellen* 'rendelkezésre bocsát' szerkezet esetében is kitűnik. Ebben a cikkben csak az elöljáró+főnév+ige típusú szerkezeteket vizsgálták, ennek megfelelően a fenti szerkezet inherens részét képező *tárgy* megtévesztő módon elmarad. Siepmann [2, 416. oldal] is hangsúlyozza: ,,az igei kollokációk és a vonzatok szorosan összefüggnek, számos ige+főnév kollokáció a vonzatok adott disztribúcióját kívánja meg . . . a vonzatuktól megfosztott ige+főnév kombinációk nem tekinthetők teljes értékű szerkezetnek".

Visszatérve a gépi fordításos példánkra, gondolhatnánk, hogy a tárgy elmaradása nem is jelent nagy problémát, mert amit az egyik nyelv tárggyal fejez ki, azt ,,nyilván" a másik is ugyanúgy tárggyal jeleníti meg. Ez azonban egyáltalán nem mindig igaz, és még kevésbé igaz az egyéb esetragokra/elöljárókra, melyek a legváltozatosabb mintázatokban felelhetnek meg egymásnak két nyelv viszonylatában.

Rendelkezésünkre áll egy korábban kifejlesztett nyelvfüggetlen lexikai kinyerő eljárás, mely képes feltérképezni egy egynyelvű korpuszban található különböző bonyolultságú igei szerkezeteket az egyszerű vonzatkeretektől (pl.: *alkalmazkodik vmihez*) a bonyolultabb, akár vonzatos, komplex igékig (pl.: *vállat von*, *örömét leli vmiben*) [6]. Az eljárást sikerrel alkalmaztunk egy egynyelvű szótár előállítása során [7].

Egy gépi fordításban közvetlenül hasznosítható kétnyelvű lexikai adatbázis vagy szótár összeállításához azonban kétnyelvű, *párhuzamos* igei szerkezetekre van szükség. Jelen dolgozatban azt vizsgáljuk, hogy hogyan adaptálható a [6]-ban leírt eljárás párhuzamos korpuszra. Azaz arra a feladatra, hogy bemenetként párhuzamos korpuszt dolgozzon fel, eredményként pedig párhuzamos igei szerkezeteket (igei szerkezeteket és a fordításukat) szolgáltasson. A távlati cél kétnyelvű lexikai adatbázis létrehozása, mely az igei szerkezetek szintjén mutatja be a két nyelv egymásnak megfelelő elemeit. Mivel az algoritmus az igei szerkezetek teljes spektrumát lefedi, azt várjuk, hogy szükség esetén képes lesz párba állítani *különböző* felépítésű szerkezeteket is: képes lesz megragadni azokat az eseteket is, amikor az egyik nyelv egyszerű igét használ ugyanarra, amit a másik nyelv komplex ige segítségével ír körül.

2. Módszer

Az eredeti lexikai kinyerő eljárás [6] tagmondatokra bontott, szintaktikailag részlegesen elemzett korpuszt vár bemenetként. A tagmondatok egy igét és annak bővítményeit kell, hogy tartalmazzák, a szintaktikai elemzésnek pedig meg kell határoznia a tagmondat igéjét, a bővítmények fejét, valamint az ige és a bővítmények közötti viszonyjelölőket. Az ilyen formátumú bemenetet feldolgozva a gyakori bővítménykereteket számbavéve az algoritmus automatikusan állítja elő a jellegzetes igei szerkezetek listáját. Az eredeti algoritmus a következőképpen működik:

1. Először is vesszük a korpuszból az összes tagmondatot. A maximum két bővítményt tartalmazó tagmondatokból váltakozva töröljük a bővítményi fejeket, így előállítjuk a tagmondatoknak megfelelő lehetséges szerkezeteket. A *Társasház jön létre,* tagmondatból kialakuló lehetséges szerkezetek: *társasház jön létre*, *vmi jön létre*, *társasház jön vmire* és *vmi jön vmire*. (Ezek közül nyilván a második a kívánt helyes szerkezet.) Erre az átalakításra azért van szükség, hogy a szerkezetlistában megjelenhessenek a szabad viszonyjelölőt (azaz esetleges vonzatot) tartalmazó szerkezetek.
2. Hossz szerint csökkenő sorba rendezzük az igei szerkezetek 1. lépés szerint kiegészített teljes listáját. Egy szerkezet hosszát a benne található esetek és fejek összesített száma adja.
3. A leghosszabbtól kezdve sorra elhagyjuk azokat a szerkezeteket, melyeknek a gyakorisága 5-nél kisebb. Az elhagyott szerkezetek gyakoriságát az első olyan rövidebb keret gyakoriságához adjuk hozzá, mely illeszkedik az eredeti keretre. (A *vmi jön létre* 3 hosszúságú keret például illeszkedik a *társasház jön létre* 4 hosszúságú keretre.) A listán még egyszer végighaladva ellenőrizzük, hogy az elhagyott szerkezetek gyakorisága mindig valóban a lehető legspecifikusabb megmaradó szerkezethez rendelődjön hozzá.
4. Végül a megmaradó szerkezetek gyakorisági érték szerint rendezett listája adja az összegyűjtött igei szerkezeteket.

3. A módszer alkalmazása párhuzamos korpuszra

Jelen munkálathoz a Dutch Parallel Corpus (Holland Párhuzamos Korpusz) [8] francia-holland részét használtuk. Ez egy könnyen hozzáférhető, morfológiailag elemzett korpusz, mely 3,2 millió holland és 3,6 millió francia tokent tartalmaz. A nyelvválasztás lehetőséget ad arra, hogy az eredetileg magyar nyelvre használt algoritmus nyelvfüggetlenségét alátámasszuk.

Az előfeldolgozás első lépéseként elvégeztük a tagmondatra bontást mindkét nyelvre. Egyszerű, szabályalapú módszerünk a következő szabályokat tartalmazta. A mondathatáron kívül tagmondathatárt jelentett a kötőszó, az alárendelt tagmondatot bevezető holland *te*, ill. francia *pour*, a vonatkozó névmás és bizonyos írásjelek (vessző, kettőspont és pontosvessző) is, amennyiben a legutóbbi tagmondathatár óta szerepelt a mondatban ige. A részleges szintaktikai elemzést szintén egyszerű szabályok használatával valósítottuk meg. A tagmondatokban

lévő főnevek (illetve a reflexív igék miatt a holland *zich* és a francia *se*) lettek a bővítményi fejek, az elöljárók pedig a viszonyjelölők. A francia *à* elöljáró + *le* névelő összevonásából keletkező *au* szócska szótövét a korpuszban lévő *au*-ról *à*-ra javítottuk, így egységesen kaptuk meg az összes *à* elöljárós bővítményt; hasonlóan jártunk el a *de* + *le* = *du* esetében is. Ha nem találtunk a fej előtt elöljárót, akkor az ige előtt alanyként, az ige után pedig tárgyként kezeltük a szóban forgó bővítményt.

Az így előállított két elemzett „félkorpuszból" a következő módon alakítottuk ki a kétnyelvű bemeneti korpuszt:

1. Az igét tartalmazó tagmondatokat fordítási egységenként sorra egymáshoz rendeltük (a fordítási egység első holland tagmondatához a megfelelő fordítási egység első francia tagmondatát stb.). Ha a fordítási egység nem azonos számú tagmondatot tartalmazott, akkor a fennmaradó(ka)t figyelmen kívül hagytuk.
2. Az egymáshoz rendelt tagmondatok holland, ill. francia igéjéből egy igepárt hoztunk létre (pl.: *gaan+aller* 'megy'), ez játssza majd az eredeti eljárás igéjének szerepét.
3. A tagmondatpárban található bővítményi csoportokat (mindkét nyelvűeket) egy halmazként soroltuk fel az igepár mellett, az egyes bővítményeket a megfelelő nyelv kódjával megjelölve.

A fenti lépések során egyfajta metakorpuszt alakítottunk tehát ki, mely párhuzamos tagmondatokból áll, a két eredeti tagmondat igéje egy metaigét alkot, a bővítmények pedig egy egyesített halmazként állnak a metaige mellett. A reprezentációt az 1. ábrán látható példa szemlélteti.

holland tagmondat: *Ze geloofde in de grote liefde*
francia tagmondat: *Elle croyait au grand amour*
magyar fordítás: 'Hitt a nagy szerelemben'

reprezentáció: geloven+croire in_{nl}:liefde à_{fr}:amour

1. ábra. Példa a kétnyelvű bemeneti korpuszból. Az igepárt '+' jel kapcsolja össze, az elöljárókat alsóindex sorolja a megfelelő nyelvhez, kettőspont után pedig az elöljáróhoz tartozó főnévi fej szerepel.

Ezek után az így kialakított kétnyelvű reprezentációra közvetlenül futtattuk az eredeti algoritmust. Mindössze két apróbb szükséges változtatást tettünk meg:

– Az algoritmus eredetileg két bővítményi pozíciót kezelt, ezt most *négyre* bővítettük, hogy hogy megkaphassuk azokat a párhuzamos szerkezeteket is, melyben mindkét nyelvben 2-2 (tehát párhuzamos szerkezetenként összesen négy) lényeges bővítmény van.

– A három és négy pozíciót tartalmazó keretek közül a vonzatos komplex ige formájúak hosszához hozzáadtunk egy 0,2-t. Így ezeknek a szerkezeteknek az esélyét megnöveltük, hogy az algoritmus 3. lépésében (104. oldal) a kiesőktől gyakoriságot örökölhessenek. E heurisztika hatására a végső listában több vonzatos komplex igét kaptunk.

4. Kiértékelés

A bemeneti kétnyelvű metakorpuszban 20-szor vagy annál többször előforduló 1356 igepárra futtattuk az algoritmust. Bár számos egy vagy két egyszerű vonzatot tartalmazó szerkezet is került az eredménylistára (l. 2. ábra), a kiértékelést mégis csak a legizgalmasabb részre, a (leggyakoribb) vonzatos komplex igékre korlátoztam.

párhuzamos szerkezet: geven+donner OBJ_{nl} aan_{nl} OBJ_{fr} à_{fr}
holland szerkezet: *geven OBJ aan*
francia szerkezet: *donner OBJ à*
magyar megfelelő: 'ad vmit vkinek'

párhuzamos szerkezet: geloven+croire in_{nl} à_{fr}
holland szerkezet: *geloven in*
francia szerkezet: *croire à*
magyar megfelelő: 'hisz vmiben'

2. ábra. Példák egyszerű vonzatot tartalmazó szerkezetekre. A párhuzamos szerkezetekből egyszerűen levezethetők a holland és francia szerkezetek, így a párhuzamos szerkezet közvetlenül megmutatja az adott igével használandó megfelelő elöljárót.

Összesen 67 olyan, legalább 15-ös gyakorisági értékkel bíró szerkezetet kaptunk, melyben vonzati pozíció és lexikálisan kötött bővítményi pozíció is volt. Az alábbi szempontok alapján fogadtam el egy párhuzamos szerkezetet helyesnek:

– Ami értelmes, az helyesnek számít, függetlenül attól, hogy idiomatikus-e a jelentése vagy sem.
– A holland *van*, ill. francia *de* általában az elemzés által egyáltalán nem kezelt birtokos szerkezetek miatt jelent meg. Ezeket nem vettük figyelembe, nem befolyásolták a szerkezetek helyességét.
– Az alany és a tárgy megállapítása nem tökéletes, ezért az alany és a tárgyat egymás helyett is elfogadtuk.
– Helyesnek fogadtuk el a szerkezetet akkor is, ha határozószó hiányzott belőle, mivel az elemzés nem kezelte a határozószókat.
– A hiányos szerkezetek nem jók, a helyességhez szükséges minden lényeges elem megléte.

1. táblázat. A kinyert 34 helyes vonzatos komplexige-szerkezet. A második és harmadik oszlopban a párhuzamos szerkezetből derivált holland, illetve francia szerkezet olvasható. A negyedik oszlopban a párhuzamos szerkezet gyakorisági értéke található. Az elöljárót a hozzá tartozó szóhoz kettőspont kapcsolja.

#	holland szerkezet	francia szerkezet	gyak	magyar megfelelő	megjegyzés
1.	*gaan om*	*agir se de*	114	'szó van vmiről'	'agir se de' 1. megfelelője
2.	*zijn OBJ*	*agir se de*	69	'vmi van'	'agir se de' 2. megfelelője
3.	*houden rekening(OBJ) met*	*tenir compte(OBJ) de*	40	'számításba vesz vmit'	met ∼ számol vmivel, tenir ∼ számon tart vmit
4.	*hebben OBJ*	*avoir besoin(OBJ) de*	39	'szükség van vmire'	holland: határozószó ('nodig') hiányzik
5.	*bestaan uit*	*composer se de*	35	'áll vmiből'	aszimmetrikus
6.	*stellen te:beschikking van*	*mettre à:disposition de*	31	'rendelkezésére bocsát'	a tárgy már nem fért bele a 4 pozícióba
7.	*spelen rol(OBJ) in*	*jouer rôle(OBJ) dans*	30	'szerepet játszik vmiben'	
8.	*bedoeld in:artikel*	*viser OBJ à:article*	30	'hivatkozik paragrafusban'	
9.	*doen beroep(OBJ) op*	*faire appel(OBJ) à*	29	'fellebbez vkihez'	
10.	*betreffen OBJ*	*agir se de*	27	'kb. illeti'	'agir se de' 3. megfelelője
11.	*zijn stad(SBJ) OBJ*	*être ville(SBJ) OBJ*	26	'a város vmilyen'	
12.	*vermelden in:artikel*	*viser OBJ à:article*	24	'említ paragrafusban'	
13.	*maken deel(OBJ) van*	*faire partie(OBJ) de*	24	'részét képezi vminek, tartozik vmihez'	
14.	*gaan over*	*agir se de*	24	'szó van vmiről'	'agir se de' 4. megfelelője
15.	*zien afbeelding(OBJ)*	*voir figurer(OBJ) de*	23	'lásd az ábrát'	
16.	*zijn van:toepassing op*	*appliquer se à*	22	'érvényes, vonatkozik vmire'	'appliquer se à' 1. megfelelője, aszimmetrikus
17.	*gelden voor*	*appliquer se à*	22	'érvényes, vonatkozik vmire'	'appliquer se à' 2. megfelelője, aszimmetrikus
18.	*nemen deel(OBJ) aan*	*participer à*	21	'részt vesz vmiben'	aszimmetrikus
19.	*richten zich tot*	*adresser se à*	19	'megcéloz, megszólít vkit'	
20.	*kennen voordeel(OBJ)*	*octroyer avantage(OBJ) de*	19	'megvan az előnye vminek'	
21.	*houden rekening(OBJ) met*	*prendre en*	19	'számításba vesz vmit'	ti. en:compte/considération
22.	*hebben betrekking(OBJ) op*	*concerner OBJ*	19	'vonatkozik vmire'	aszimmetrikus
23.	*zijn op:zoek naar*	*être à:recherche de*	18	'keres vmit'	
24.	*heten*	*appeler se OBJ*	18	'hívják vhogy'	
25.	*hebben effect(OBJ) op*	*avoir effet(OBJ) sur*	18	'(vmilyen) hatása van vmire'	
26.	*zijn in:België*	*être en:Belgique de*	17	'van Belgiumban'	
27.	*vergaderen*	*réunir se de*	17	'találkozót tart, összeül'	
28.	*zijn OBJ*	*etre OBJ à:foi*	16	'egyszerre van'	'à la fois' = ugyanakkor + holland határozószó
29.	*stoppen*	*arreter se de*	16	'befejeződik'	
30.	*liggen aan:basis van*	*etre à:base de*	16	'vminek az alapja'	
31.	*branden*	*allumer se de*	16	'ég (pl. lámpa)'	
32.	*bedragen euro(OBJ)*	*élever se à*	16	'(vmennyi euró) össszeget tesz ki'	aszimmetrikus, hiányzik a francia 'euro'
33.	*zijn OBJ*	*faire objet(OBJ) de*	15	'vmi alanya lesz' ???	
34.	*spelen rol(OBJ)*	*jouer rôle(OBJ) de*	15	'szerepet játszik'	'vmiben' nélküli változat

A fenti szempontok miatt 9 szerkezet egy másik szerkezettel egybeesett, ezeket kizártuk az értékelésből, sem helyesnek, sem helytelennek nem számítottuk. A megmaradó 58 szerkezetből a kiértékelés során 34 bizonyult helyesnek, ez *58,6 százalékos* pontosságot jelent. Ez természetesen jócskán elmarad az eredeti cikkben közölt, egynyelvű, magyar korpuszon mért 94 százalékos pontossági értéktől. Jelen feladat nyilvánvalóan jóval nehezebb: sokkal több elemet kell helyesen megtalálni, hogy a kapott párhuzamos szerkezet valóban teljes legyen. A 34 helyes vonzatos komplexige-szerkezetet az 1. táblázat tartalmazza.

5. Példák

A bevezető végén elővételeztük, hogy az algoritmusunk várhatóan leghasznosabb tulajdonsága az lesz, hogy olyan párhuzamos szerkezetek felfedezésére is képes, ahol a két nyelv teljesen más felépítésű szerkezetet használ az adott jelentés kifejezésére. Ezeket a párhuzamos szerkezeteket *aszimmetrikus*nak nevezzük. Gyengén vagy ,,tartalmilag" aszimmetrikus egy párhuzamos szerkezet, ha ugyanannyi szabad, illetve lexikálisan kötött bővítménye van, de a bővítmények nem az alapértelmezett módon felelnek meg egymásnak: tárgynak nem tárgy felel meg, vagy a kötött szavaknak, illetve a viszonyjelölőknek nem a szokásos fordítása szerepel. Erősen vagy ,,formailag" aszimmetrikus egy párhuzamos szerkezet, ha a bővítmények közvetlenül nem feleltethetők meg egymásnak, vagy a bővítmények száma nem is egyezik a két nyelvben. Az 1. táblázatban aszimmetrikusként megjelölt szerkezetek közül a legérdekesebb a következő három:

- A 18. sorszámú szerkezet klasszikus példája az egyszerű és komplex ige megfelelésének: a *részt vesz* fogalmát a holland nyelv a magyarhoz hasonlóan komplex igével (*nemen deel(OBJ)*) fejezi ki, a francia pedig egy szóval (*participer*).
- A 22. sorszámú szerkezet aszimmetriáját az (is) okozza, hogy a francia tárgy a hollandban nem tárgynak, hanem *op* elöljárós bővítménynek felel meg.
- A legbonyolultabb a 16. sorszámú szerkezet: itt a francia részen reflexív igével (*appliquer se*) találkozunk, a hollandban pedig egy létigés komplex igével (*zijn van:toepassing*).

Itt térhetünk ki annak a felvetődő kérdésnek a megválaszolására, hogy eredetileg miért nem úgy fogtunk neki a feladatnak, hogy külön-külön ellőállítottuk volna a holland és francia szerkezeteket, majd a két szerkezettárat illesztettük volna össze. A válasz az, hogy azért, mert így megkaphatjuk azokat a párhuzamos szerkezeteket is, amelyek két oldala formailag egyáltalán nem hasonlít egymásra.

Az eredmények jól mutatják az ismert tényt, hogy a különböző nyelvek egyes nyelvi elemei csak nagyjából felelnek meg egymásnak: sokszor van példa arra, hogy az egymás fordításának vélt szavak csak bizonyos környezetben fordításai egymásnak, vagy bizonyos környezetben nem fordításai egymásnak. Másképp fogalmazva a nyelvi elemek (például igék vagy elöljárók), a kifejezések különböző részhalmazait fedik le, és két nyelv viszonylatában ezek a részhalmazok szinte

soha nem esnek pontosan egybe, az átfedés mértéke széles határok között változik. Mikor egy párhuzamos szerkezetben egy tartalmas szónak nem a szokásos fordítása van jelen, máris egy gyengén aszimmetrikus szerkezettel van dolgunk.

A párhuzamos szerkezetek szépen megadják az igék egy-egy „jelentését", pontosabban azt, hogy adott környezetben, az adott elemek mellé éppen melyik ige illik. A szerkezet többi része sok esetben „szó szerinti" fordítás, és pontosan az ige az, amely kifejezésről kifejezésre más-más, nem kikövetkeztethető, megtanulandó, idiomatikus. Így van ez a 9. és a 13. szerkezet (l.: 1. táblázat) esetében, mikor a 'csinál' jelentésű francia *faire* az egyik kifejezésben a hasonló jelentésű holland *doen*-nal áll párban, máskor pedig a szintén hasonló jelentésű *maken*-nel, de nem felcserélhető módon.

Hasonlóan viselkednek az elöljárók is, gyakran kevéssé megjósolható módon. A nagyjából *-on/-en/-ön* vagy *-ra/re* szerepű elöljárók közül valamikor az *op–à* (16. szerkezet), máskor pedig az *aan–à* (18. szerkezet) áll párban, ugyanakkor az *op*-nak a *sur* is megfelelhet (25. szerkezet).

6. Összefoglalás

Az ismertetett módszer korpuszvezérelt módon, kétnyelvű igei szerkezetek hasznos gyűjteményét képes előállítani. Lényeges tulajdonsága, hogy képes felfedezni a formailag egymásra nem hasonlító, de egymásnak megfelelő, egymás fordításaiként kezelendő igei szerkezeteket. Nehéz feladat a párhuzamos szerkezetekben lévő számos elem mindegyikét megtalálni, ezért gyakran előfordul, hogy a kapott szerkezetek hiányosak. Ilyen esetben kézi utószerkesztéssel lehet javítani a hibákat.

A nyelvenkénti 3-3,5 millió szavas korpusz ilyen feladatra kicsinek számít, ezért viszonylag alacsony a kapott szerkezetek száma. A párhuzamos korpuszok előállítási költsége magas, ezért a közeljövőben maximum ennél egy nagyságrenddel nagyobb párhuzamos korpuszokra számíthatunk. Ezek használata azonban már jelentősen növelhetné a kinyerhető párhuzamos szerkezetek mennyiségét.

Amint a fentiekben láttuk, rendre egyszerű közelítő módszereket alkalmaztunk az előkészítő, elemző lépések során. Az e lépések során előforduló különféle hibáktól, hiányosságoktól függetlenül egyértelművé vált a módszer képessége az egymásnak megfelelő igei szerkezetek közvetlen megragadására. Az elemzési lépések fejlesztése nagy mértékben javíthatna a végső eredmény minőségén, de az a jelen dolgozatból így is látszik, hogy maga az algoritmus megfelel a kívánt célnak.

Ha egy párhuzamos szerkezet két oldalán 1-1 vonzati pozíció van, akkor azok a legtöbb esetben egymás megfelelői. Ritkán előfordulhat több ilyen pozíció (pl.: *őrizetbe vesz vkit vmi miatt*), ekkor egy külön módszerrel kell meghatározni, hogy melyik vonzati pozíció melyiknek felelhet meg, például az előforduló tartalmas szavak élősége/élettelensége alapján. Ennek kidolgozása a jövő feladata, mint ahogy a további, egyszerűbb típusokra kiterjedő kiértékelés elvégzése is.

Hivatkozások

1. Evert, S.: The Statistics of Word Cooccurrences: Word Pairs and Collocations. PhD thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart (2005)
2. Siepmann, D.: Collocation, colligation and encoding dictionaries. Part I: Lexicological aspects. International Journal of Lexicography **18**(4) (2005) 409–444
3. Pecina, P.: A machine learning approach to multiword expression extraction. In: Proceedings of the LREC2008 workshop: Towards a Shared Task for Multiword Expressions, Marrakech, Morocco (2008) 54–57
4. Seretan, V.: Collocation extraction based on syntactic parsing. PhD thesis, University of Geneva (2008)
5. Evert, S., Krenn, B.: Methods for the qualitative evaluation of lexical association measures. In: Proceedings of the 39th Meeting of the Association for Computational Linguistics, Toulouse, France (2001)
6. Sass, B.: A unified method for extracting simple and multiword verbs with valence information and application for Hungarian. In: Proceedings of RANLP 2009, Borovets, Bulgaria (2009) 399–403
7. Sass, B., Pajzs, J.: FDVC – creating a corpus-driven frequency dictionary of verb phrase constructions. In: eLexicography in the 21st century: New challenges, new applications. Proceedings of eLex 2009, Cahiers du CENTAL 7. Presses universitaires de Louvain, Louvain-la-Neuve, Belgium (2010) 263–272
8. Macken, L., Trushkina, J., Paulussen, H., Rura, L., Desmet, P., Vandeweghe, W.: Dutch Parallel Corpus. A multilingual annotated corpus. In: Proceedings of Corpus Linguistics 2007, Birmingham, United Kingdom (2007)